Kernelized Supervised Dictionary Learning

Mehrdad J. Gangeh*, Ali Ghodsi, and Mohamed S. Kamel

***Abstract*—In this paper, we propose supervised dictionary learning (SDL) by incorporating information on class labels into the learning of dictionary. To this end, we propose to learn the dictionary in a space where the dependency between the signals and their corresponding labels is maximized. To maximize this dependency, the recently introduced Hilbert Schmidt independence criterion (HSIC) is used. One of the main advantages of this novel approach for SDL is that it can be easily kernelized by incorporating a kernel, particularly a data-derived kernel such as normalized compression distance, into the formulation. The learned dictionary is compact and the proposed approach is fast. We show that it outperforms other unsupervised and supervised dictionary learning approaches in the literature on real-world data.**

***Index Terms*—Pattern recognition and classification, classification methods, non-parametric methods, dictionary learning, HSIC, supervised learning.**

I. Introduction

DICTIONARY learning and sparse representation (DLSR) are two closely related topics that have roots to the decomposition of signals to some predefined bases such as Fourier transform. However, what make DLSR distinct from the representation using predefined bases are that first, the bases are learned here from data and second, only few components in the dictionary are needed to represent data (sparse representation). This latter attribute can be also seen in the decomposition of signals using some predefined bases such as wavelets [1].

The concept of dictionary learning and sparse representation was originated from different communities to solve different problems, which are given different names. Some of them are: sparse coding (SC), which was originated by neurologists as a model for simple cells in mammalian primary visual cortex [2]; independent component analysis (ICA), which was originated by researchers in signal processing to estimate the underlying hidden components of multivariate statistical data (refer to [3] for a review of ICA); least absolute shrinkage and selection operator (*lasso*), which was originated by statisticians to find linear regression models when there are many more predictors than samples, where some constraint has to be considered to fit the model. In the *lasso*, one of the constraints introduced by Tibshirani was the ℓ_1 norm that led to sparse coefficients in the linear regression model [4]. Another technique which also leads to DLSR is Nonnegative matrix factorization (NNMF), which aimed to decompose a matrix to two nonnegative matrices, one of which can be considered as the dictionary and the other as the coefficients [5]. In NNMF, usually both the dictionary and coefficients are sparse [5], [6]. This list is not complete and there are variants for each of the above techniques such as blind source separation (BSS) [7], compressed sensing [8], basis pursuit (BP) [9], and orthogonal matching pursuit (OMP) [10], [11]. It is beyond the scope of this paper to include the description of all these techniques (interested readers can refer to [12]–[14] for a review on dictionary learning and sparse representation).

The main results of all these research works is that a class of signals with sparse nature, such as images of natural scenes, can be represented using some *primitive elements* that form a dictionary, and that each signal in this class, can be represented by using only few elements in the dictionary (sparse representation). In fact, there are, at least, two ways in the literature to exploit sparsity [15]: first, using a linear/nonlinear combination of some predefined bases, e.g., wavelets [1]. Second, by using primitive elements in a learned dictionary, such as techniques employed in SC or ICA. This latter approach is our focus in this paper and has led to state-of-the-art results in various applications such as texture classification [16]–[18], face recognition [19]–[21], image denoising [22], [23], etc.

We may categorize the various dictionary learning with sparse representation approaches proposed in the literature in different ways. One way is based on whether the dictionary is consisting of predefined or learned bases as stated above. Another way is based on the model used to learn the dictionary and coefficients. These models can be *generative* as what is used in original formulation of SC [2], ICA [3], and NNMF [5]; *reconstructive* as in the *lasso* [4]; or discriminative such as SDL-D (supervised dictionary learning-discriminative) in [15]. The two former approaches do not consider the class labels in building the dictionary while the last one, i.e., discriminative one does. In other words, we state that dictionary learning can be performed unsupervised or supervised, with the difference that in the latter, the class labels in the training set are used to build a more discriminative dictionary for the particular classification task in hand.

In this paper, we propose a novel supervised dictionary learning (SDL) by incorporating information on class labels into the learning of dictionary. The dictionary is learned in a space where the dependency between the data and their corresponding labels is maximized. We propose to maximize this dependency by using the recently introduced Hilbert Schmidt independence criterion (HSIC) [24], [25]. The dictionary is then learned in this new space. Although, supervised dictionary learning has been proposed by others as will be reviewed in next section, this work is different from others in following

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Canada Graduate Scholarship (CGS D3-378361-2009).

M. J. Gangeh and M. S. Kamel are with the Center for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ONT. N2L 3G1, Canada e-mail: {mgangeh,mkamel}@pami.uwaterloo.ca.

A. Ghodsi is with the Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ONT. N2L 3G1, Canada e-mail: aghodsib@uwaterloo.ca.

aspects:

1) The formulation is simple and straightforward.
2) The proposed approach introduces a closed form formulation for the computation of dictionary. This is different from other approaches, in which the computation of dictionary and sparse coefficients has to be iteratively (and often alternatively) performed which causes high computational load.
3) The approach is very efficient in terms of dictionary size (compact dictionary). Our results show that the proposed dictionary can particularly produce significantly better results than other supervised dictionary methods at small dictionary sizes.
4) The proposed approach can be easily kernelized by incorporating a kernel into the formulation. Data dependent kernels based on, e.g., normalized compression distance (NCD) [26], [27] can be used in this kernelized SDL to further improve the discrimination power of the designed system. To our best of knowledge, no other kernelized SDL approach has been proposed in the literature yet and none of the proposed SDLs in the literature can be kernelized in a straightforward way.

The organization of the rest of the paper is as follows: in Section II, we review the current SDL approaches in the literature and their shortcomings. Then we review the mathematical background and the formulation for proposed approach in Section III. The experimental setup and results are presented in Sections IV, followed by discussion and conclusion in Section V.

II. Background and Related Work

In this section, we provide an overview on the dictionary learning and sparse representation and a brief review of recent attempts on making the approach more suitable for classification tasks.

A. Dictionary Learning and Sparse Representation

Considering a finite training set of signals $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{p\times n}$, where p is the dimensionality and n is the number of data samples, according to the classical dictionary learning and sparse representation (DLSR) techniques (refer to [12] and [13] for a recent review on this topic), these signals can be represented by a linear decomposition over few dictionary atoms by minimizing a loss function as given below

$$L(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} l(\mathbf{x}_i, \mathbf{D}, \boldsymbol{\alpha}), \tag{1}$$

where $\mathbf{D} \in \mathbb{R}^{p\times k}$ is the dictionary of k atoms, and $\boldsymbol{\alpha} \in \mathbb{R}^{k\times n}$ are the coefficients.

This loss function can be defined in various ways based on the application in hand. However, what is common in DLSR literature is to define the loss function L as the reconstruction error in a mean-squared sense with a sparsity inducing function ψ as a regularization penalty to ensure the sparsity of coefficients. Hence, (1) can be written as

$$L(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) = \min_{\mathbf{D},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\boldsymbol{\alpha}\|_{\mathrm{F}}^2 + \lambda\psi(\boldsymbol{\alpha}), \tag{2}$$

where subscript F indicates Frobenius norm and λ is the regularization parameter that affects the number of nonzero coefficients.

An intuitive measure of sparsity is ℓ_0 norm, which indicates the number of nonzero elements in a vector[1]. However, the optimization problem obtained from replacing sparsity inducing function ψ in (2) with ℓ_0 is nonconvex and the problem is NP-hard (refer to [13] for a recent comprehensive discussion on this issue). Two main proposed (approximate) solutions to overcome this problem is first, based on greedy algorithms, such as well-known orthogonal matching pursuit (OMP) [10], [11], [13]. Second, by approximating highly discontinuous ℓ_0 norm by a continuous functions such as the ℓ_1 norm. This leads to an approach, which is widely known in literature as *lasso* [4] or *basis pursuit* (BP) [9] and (2) converts to

$$L(\mathbf{X}, \mathbf{D}, \boldsymbol{\alpha}) = \min_{\mathbf{D},\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{X} - \mathbf{D}\boldsymbol{\alpha}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{\alpha}\|_1. \tag{3}$$

In (3), the main optimization goal for computation of dictionary and sparse coefficients is minimizing the reconstruction error in mean-squared sense. While this works well in applications where the primary goal is to reconstruct signals as accurate as passible such as in denoising, image inpainting, and coding, it is not the ultimate goal in classification tasks [28] as discriminating signals is more important here. Hence, recently, there has been several attempts to include category information in computing dictionary, coefficients, or both. In following subsection, we will provide a brief overview of proposed supervised dictionary learning approaches in the literature. To this end, we will try to categorize the proposed approaches into five different categories, while we admit that this taxonomy of approaches is not unique and it can be done differently.

B. Supervised Dictionary Learning in Literature

As mentioned in previous subsection, (3) provides a reconstructive formulation for computing the dictionary and sparse coefficients given a set of data samples. Although the problem is not convex on both dictionary $\mathbf{D}$ and coefficients $\boldsymbol{\alpha}$, this optimization problem is convex if it is solved iteratively and alternatively on these two unknowns. Several fast algorithms have been recently proposed for this purpose such as K-SVD [29], online learning [30], and cyclic coordinate descent [31]. However, none of these approaches take into account the category information for learning the dictionary or coefficients.

The first and simplest approach to include category information in DLSR is computing one dictionary per class, i.e., using the training samples in each class to compute part of dictionary and then compose all these partial dictionaries into one. Perhaps the earliest work in this direction is so called texton-based approach [18], [32], [33]. In this approach, *k*-means is applied to the training samples in each class and the *k* cluster centers computed are considered as the dictionary for this class. These partial dictionaries are eventually composed into one dictionary. In [20], the training samples are used

[1] ℓ_0 norm of vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_0 = \#\{i : x_i \neq 0\}$.

as the dictionary in face recognition and hence, basically, it falls in the same category as training one dictionary per class. However, no actual training is performed here and the whole training samples are used directly in the dictionary. Using the training samples as dictionary yields a very large and possibly inefficient dictionary due to noisy training instances. To obtain smaller dictionary, Yang et al. propose to learn a smaller dictionary per class called *metaface* (proposed approach was in face recognition application but it is general and can be used in any application) and then compose them into one dictionary [34]. One major drawback of this approach is that the training samples in one class are used for computing the atoms in the dictionary irrespective of the training samples form other classes. This means that if training samples across classes have some common properties, these shared properties cannot be learned in common in the dictionary. Ramirez et al. propose to overcome this problem by including an incoherence term to (3) to encourage independency of dictionaries from different classes while still allowing for different classes to share features [35]. The main drawback of all approaches in this first category of SDL is that they may lead to very large dictionary as the size of composed dictionary grows linearly with the number of classes.

The second category of SDL approaches learn a (very) large dictionary unsupervised in the beginning. Then merge the atoms in the dictionary by optimizing an objective function that takes into account the category information. One major work in literature in this direction is based on information bottleneck that iteratively merges two dictionary atoms that cause the smallest decrease in the mutual information between dictionary atoms and class labels [36]. Another major work is based on merging two dictionary atoms that minimizes the loss of mutual information between histogram of dictionary atoms, over signal constitutes (e.g., image patches), and class labels [37]. One main drawback of this category of SDL is that reduced dictionary obtained usually performs at most the same as original one. Hence, since the initial dictionary is learned unsupervised, although due to its large size it includes almost all possible atoms that helps to improve the performance of classification task, the consecutive pruning stage is inefficient in terms of computational load and it can be significantly improved by finding a discriminative dictionary from the beginning.

The third category of SDL, which is based on several research works published in [15], [38]–[42] can be considered a major leap in SDL. In this category, the classifier parameters and dictionary are learned in a joint optimization problem. Although this idea is more sophisticated than the previous two, its major disadvantage is that the optimization problem is nonconvex and complex. If it is done alternatively between dictionary learning and classifier parameters learning, it is quite likely that they stuck in local minima. On the other hand, due to the complexity of the problem, except for bilinear classifier in [15], other papers only consider linear classifiers which is usually too simple to solve difficult problems and can only be successful in simple classification tasks as shown in [15]. In [39], Zhang and Li propose a technique called discriminative K-SVD (DK-SVD), which is truly jointly learn the classifier parameters and dictionary without alternating between these two steps. This prevents the possibility of getting stuck in local minima. However, only linear classifiers are considered in DK-SVD that may lead to poor performance in difficult classification tasks. Another major problem with the approaches in this category of SDL is that there exist many parameters involved in the formulation, which are hard and time consuming to be tuned (see for example [15], [42]).

The fourth category of SDL approaches include the category information into the learning of dictionary. This is done, for example, by minimizing the information loss due to predicting labels from supervised dictionary learned instead of original training data samples (this approach is known as *info-loss* in SDL literature) [43] or by deploying extremely randomized decision forests [44] (this latter approach can also fall in the second category of SDLs as it seems that it starts from a very large dictionary using random forests and try to prune it later to conclude a smaller dictionary). The info-loss approach has this major drawback that it may also stuck in local minima (the same as previous category of SDL) and the optimization has to be done iteratively and alternatively on two updates as there is no closed form solution for the approach.

The fifth category of SDLs, include class category in learning the coefficients [28] or in learning both dictionary and coefficients [21], [45]. Supervised coefficient learning in all these papers [21], [28], [45] has been performed more or less the same using Fisher discrimination criterion [46], i.e., by minimizing the within-class covariance of coefficients and at the same time maximizing their between-class covariance. As for the dictionary, while [28] uses predefined bases, [21] proposes a discriminative fidelity term that encourages learning dictionary atoms of one class from the training samples of the same class and at the same time penalizes their learning by the training samples from other classes. The joint optimization problem due to Fisher discrimination criterion on the coefficients and discriminative fidelity term on the dictionary proposed in [21] is not convex and has to be solved alternatively and iteratively between these two terms until it converges. However, there is no guarantee in this approach to find the global minimum. Also, it is not clear whether the improvement obtained in classification by including Fisher discriminant criterion on coefficients justifies the additional computation load imposed on the learning as there is no comparison provided in [21] on the classification with and without including supervision on coefficients.

In next section, we explain the mathematical formulation for our proposed approach, which we believe, belongs to the fourth category of SDLs explained above, i.e., including category information to learn a supervised dictionary.

III. Methods

To incorporate the category information into the dictionary learning, we propose to decompose the signals using some learned bases that represent them in a space where the dependency between the signals and their corresponding class labels is maximized. To this end, we need a(n) (in)dependency test measure between two random variables. Here, we propose

to use Hilbert-Schmidt independence criterion (HSIC) as the (in)dependency measure. In this section, we first describe HSIC and then provide the formulation for our proposed supervised dictionary learning (SDL) approach. Subsequently, kernelized SDL is formulated that enables embedding kernels including data-dependent ones into the proposed SDL. This can significantly improve the discrimination power of designed dictionary, which is essential in difficult classification tasks as will be shown in our experiments in Subsection IV-E later.

A. Hilbert Schmidt Independence Criterion

There are several techniques in literature to measure the (in)dependence of random variables such as mutual information [47] and Kullback-Leibler (KL) divergence [48]. In addition to these measures, there has been recently great interest in measuring (in)dependency using criteria based on functions in reproducing kernel Hilbert spaces (RKHSs). Bach and Jordan were those who first accomplished this by introducing kernel dependence functionals that significantly outperformed alternative approaches [49]. Later, Gretton et al. proposed another kernel-based approach called Hilbert-Schmidt independence criterion (HSIC) to measure the (in)dependence of two random variables $\mathcal{X}$ and $\mathcal{Y}$ [24]. Since its introduction, the HSIC has been used in many applications including feature selection [50], independent component analysis [51], and sorting/matching [52].

One can derive HSIC as a measure of (in)dependence between two random variables $\mathcal{X}$ and $\mathcal{Y}$ using two different approaches: first by computing Hilbert-Schmidt norm of the cross-covariance operators in RKHSs as shown in [24], [25]; second, by computing maximum mean discrepancy (MMD) of two distributions mapped to a high dimensional space (i.e., computed in RKHSs) [53], [54]. We believe that this latter approach is more straightforward and hence, use it to describe HSIC.

Let $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1,), ..., (\mathbf{x}_n, \mathbf{y}_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$ be n independent observations drawn from $p := P_{\mathcal{X}\times\mathcal{Y}}$. To investigate whether $\mathcal{X}$ and $\mathcal{Y}$ are independent we need to determine whether distribution p factorizes, i.e., whether p is the same as $q := P_{\mathcal{X}} \times P_{\mathcal{Y}}$.

The mean of distributions are defined as follows

$$\mu[P_{\mathcal{X}\times\mathcal{Y}}] := \mathbf{E}_{xy}[v((x,y),.)] \quad (4)$$

$$\mu[P_{\mathcal{X}} \times P_{\mathcal{Y}}] := \mathbf{E}_x\mathbf{E}_y[v((x,y),.)] \quad (5)$$

where kernel $v((x,y),(x',y'))$ is defined in RKHS over $\mathcal{X}\times\mathcal{Y}$. By computing the mean of distributions p and q in RKHS, we effectively take into account higher order statistics than the first order by mapping these distributions to a high dimensional feature space. Hence, we can use $\text{MMD}(p,q) := \|\mu[P_{\mathcal{X}\times\mathcal{Y}}] - \mu[P_{\mathcal{X}} \times P_{\mathcal{Y}}]\|_2$ as a measure of (in)dependence of random variables $\mathcal{X}$ and $\mathcal{Y}$. The higher the value of MMD, the closer two distributions p and q and hence, the more dependent random variables $\mathcal{X}$ and $\mathcal{Y}$.

Now suppose that $v((x,y),(x',y')) = k(x,x')l(y,y')$, i.e., the RKHS is a direct product of $\mathcal{H}\otimes\mathcal{G}$ of the RKHSs on $\mathcal{X}$ and $\mathcal{Y}$. Then $\text{MMD}(p,q)$ can be written as

$$\begin{aligned}\text{MMD}^2(p,q) &= \|\mathbf{E}_{xy}[k(x,.)l(y,.)] \\ &\quad - \mathbf{E}_x[k(x,.)]\mathbf{E}_y[l(y,.)]\|_2^2 \\ &= \mathbf{E}_{xy}\mathbf{E}_{x'y'}[k(x,x')l(y,y')] \\ &\quad - 2\mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'y'}[k(x,x')l(y,y')] \\ &\quad + \mathbf{E}_x\mathbf{E}_y\mathbf{E}_{x'}\mathbf{E}_{y'}[k(x,x')l(y,y')]. \quad (6)\end{aligned}$$

This is exactly the HSIC and equivalent to the Hilbert-Schmidt norm of the cross-covariance operator in RKHSs [24].

For practical purposes, HSIC has to be estimated using a finite number of data samples. Considering $\mathcal{Z} := \{(\mathbf{x}_1, \mathbf{y}_1,), ..., (\mathbf{x}_n, \mathbf{y}_n)\} \subseteq \mathcal{X} \times \mathcal{Y}$ as n independent observations drawn from $p := P_{\mathcal{X}\times\mathcal{Y}}$, an empirical estimate of HSIC is defined as follows [24]

$$\text{HSIC}(\mathcal{Z}) = \frac{1}{(n-1)^2}\text{tr}(\mathbf{KHLH}), \quad (7)$$

where tr is the trace operator, $\mathbf{H}, \mathbf{K}, \mathbf{L} \in \mathbb{R}^{n\times n}, K_{i,j} = k(x_i, x_j), L_{i,j} = l(y_i, y_j)$, and $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{e}\mathbf{e}^\top$ ($\mathbf{I}$ is the identity matrix, and $\mathbf{e}$ is a vector of n ones, and hence, $\mathbf{H}$ is the centering matrix). It is important to notice that according to (7), to maximize the dependency between two random variables $\mathcal{X}$ and $\mathcal{Y}$, the empirical estimate of HSIC, i.e., $\text{tr}(\mathbf{KHLH})$ should be maximized.

B. Proposed Supervised Dictionary Learning

To formulate our proposed SDL, we start from the reconstruction error given in (3). Let we have a finite training set of n data points, each of which consisting of p features, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{p\times n}$. We further assume that features in data samples are centered, i.e., their mean is removed and hence, each row of $\mathbf{X}$ sums to zero. We address the problem of finding a linear decomposition of data $\mathbf{X} \in \mathbb{R}^{p\times n}$ using some bases $\mathbf{U} \in \mathbb{R}^{p\times k}$ such that the reconstruction error (in mean-squared sense) is minimum, i.e.,

$$\min_{\mathbf{U}, V_i} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{U}V_i\|_2^2, \quad (8)$$

where V_i is the vector of k reconstruction coefficients in the subspace defined by $\mathbf{U}^\top\mathbf{X}$. We can rewrite (8) in matrix form as follows

$$\min_{\mathbf{U}, \mathbf{V}} \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_{\text{F}}^2, \quad (9)$$

where $\mathbf{V} \in \mathbb{R}^{k\times n}$ is the matrix of coefficients. Since both $\mathbf{U}$ and $\mathbf{V}$ are unknown, this problem is ill-posed and does not have unique solution unless we impose some constraints on the bases $\mathbf{U}$. If we, for example, assume that the bases are orthonormal, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, (9) can be written as a constrained optimization problem as follows

$$\begin{aligned}\min_{\mathbf{U}, \mathbf{V}} \quad & \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_{\text{F}}^2. \\ \text{s.t.} \quad & \mathbf{U}^\top\mathbf{U} = \mathbf{I}\end{aligned} \quad (10)$$

To further investigate the optimization problem in (10), we assume that the matrix $\mathbf{U}$ is fixed and find the optimum matrix

of coefficients $\mathbf{V}$ in terms of $\mathbf{X}$ and $\mathbf{U}$ by taking the derivative of the objective function given in (10) in respect to $\mathbf{V}$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{V}} \|\mathbf{X}-\mathbf{U}\mathbf{V}\|_{\mathrm{F}}^2 &= \frac{\partial}{\partial \mathbf{V}}\mathrm{tr}[(\mathbf{X}-\mathbf{U}\mathbf{V})^\top(\mathbf{X}-\mathbf{U}\mathbf{V})] \\ &= \frac{\partial}{\partial \mathbf{V}}[\mathrm{tr}(\mathbf{X}^\top\mathbf{X}) - 2\mathrm{tr}(\mathbf{X}^\top\mathbf{U}\mathbf{V}) \\ &\quad + \mathrm{tr}(\mathbf{V}^\top\mathbf{U}^\top\mathbf{U}\mathbf{V})] \\ &= -2\mathbf{U}^\top\mathbf{X} + 2\mathbf{U}^\top\mathbf{U}\mathbf{V}\end{aligned}$$

Equating the above derivative to zero and knowing that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, we obtain

$$\mathbf{V} = \mathbf{U}^\top\mathbf{X}. \tag{11}$$

By plugging $\mathbf{V}$ found in (11) into objective function of (10) we obtain

$$\begin{aligned}\min_{\mathbf{U}} \left\|\mathbf{X}-\mathbf{U}\mathbf{U}^\top\mathbf{X}\right\|_{\mathrm{F}}^2 &= \min_{\mathbf{U}} \mathrm{tr}[(\mathbf{X}-\mathbf{U}\mathbf{U}^\top\mathbf{X})^\top(\mathbf{X}-\mathbf{U}\mathbf{U}^\top\mathbf{X})] \\ &= \min_{\mathbf{U}} [\mathrm{tr}(\mathbf{X}^\top\mathbf{X}) - 2\mathrm{tr}(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}) \\ &\quad + \mathrm{tr}(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X})] \\ &= \max_{\mathbf{U}} \mathrm{tr}(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}) \\ &= \max_{\mathbf{U}} \mathrm{tr}((\mathbf{U}^\top\mathbf{X})^\top\mathbf{U}^\top\mathbf{X})\end{aligned}$$

Let $\mathbf{K} = (\mathbf{U}^\top\mathbf{X})^\top\mathbf{U}^\top\mathbf{X}$, which is a linear kernel on the transformed data in the subspace $\mathbf{U}^\top\mathbf{X}$, recalling that the features are centered in the original space, we can write

$$\max_{\mathbf{U}} \mathrm{tr}((\mathbf{U}^\top\mathbf{X})^\top\mathbf{U}^\top\mathbf{X}) = \max_{\mathbf{U}} \mathrm{tr}(\mathbf{K}\mathbf{H}\mathbf{I}\mathbf{H}), \tag{12}$$

where $\mathbf{H}$ and $\mathbf{I}$ are the centering and identity matrices, respectively.

By recalling the empirical HSIC given in (7), the main conclusion from (12) is that the bases $\mathbf{U}$ represents the centered data[2] $\mathbf{X}\mathbf{H}$ in a space where each data sample has the maximum dependency to itself. We know that these bases are the principal components of the signal $\mathbf{X}$ that represent the data in an uncorrelated space. In other words, we have shown that the optimization problem in (10) is equivalent to

$$\begin{aligned}\max_{\mathbf{U}} \quad & \mathrm{tr}(\mathbf{U}^\top\mathbf{X}\mathbf{H}\mathbf{I}\mathbf{H}\mathbf{X}^\top\mathbf{U}) \\ \text{s.t.} \quad & \mathbf{U}^\top\mathbf{U} = \mathbf{I}\end{aligned} \tag{13}$$

whose solution is the top eigenvectors of $\mathbf{\Phi} = \mathbf{X}\mathbf{H}\mathbf{I}\mathbf{H}\mathbf{X}^\top$, where $\mathbf{X}\mathbf{H}\mathbf{I}\mathbf{H}\mathbf{X}^\top$ is the covariance matrix of $\mathbf{X}$.

To summarize, we found out in previous paragraphs that the linear decomposition of signals that minimizes the reconstruction error in mean-squared sense, represents the data in an uncorrelated space. However, as mentioned before, although minimization of reconstruction error is the ultimate goal in applications such as denoising and coding, in classification tasks, main goal is maximum discrimination of classes. Hence, we are looking for a decomposition that represents the data in a space where the decomposed data have maximum dependency with their labels. To this end, we propose the new optimization problem as follows

$$\begin{aligned}\max_{\mathbf{U}} \quad & \mathrm{tr}(\mathbf{U}^\top\mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^\top\mathbf{U}) \\ \text{s.t.} \quad & \mathbf{U}^\top\mathbf{U} = \mathbf{I}\end{aligned} \tag{14}$$

where $\mathbf{L}$ is a linear kernel on the labels $\mathbf{Y}$, i.e., $\mathbf{Y}\mathbf{Y}^\top$. Similar to the previous case, the solution for the optimization problem given in (14) is top eigenvectors of $\mathbf{\Phi} = \mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^\top$. These eigenvectors compose the supervised dictionary to be learned. This dictionary spans the space where the dependency between data $\mathbf{X}$ and corresponding labels $\mathbf{Y}$ is maximized. The coefficients can be computed in this space using the *lasso* as given in (3). The optimization problem given in (14), compromises the reconstruction error to achieve a better discrimination power. In conclusion, we propose our supervised dictionary learning as given in Algorithm 1.

One important advantage of proposed approach in Algorithm 1 is that the dictionary can be computed in closed form. Besides, learning dictionary and coefficients is performed separately and we do not need to learn these two iteratively and alternatively as is common in most of supervised dictionary learning approaches in the literature (refer to Subsection II-B).

Algorithm 1 Supervised Dictionary Learning

Input: Training data, $\mathbf{X}_{\text{tr}}$, test data, $\mathbf{X}_{\text{ts}}$, kernel matrix of labels $\mathbf{L}$, training data size, n, size of dictionary, k.
Output: Dictionary, $\mathbf{D}$, coefficients for training and test data, $\boldsymbol{\alpha}_{\text{tr}}$ and $\boldsymbol{\alpha}_{\text{ts}}$.

1: $\mathbf{H} \leftarrow \mathbf{I} - n^{-1}\mathbf{e}\mathbf{e}^\top$
2: $\mathbf{\Phi} \leftarrow \mathbf{X}\mathbf{H}\mathbf{L}\mathbf{H}\mathbf{X}^\top$
3: **Compute Dictionary:** $\mathbf{D} \leftarrow$ eigenvectors of $\mathbf{\Phi}$ corresponding to top k eigenvalues.
4: **Compute Training Coefficients:** $\mathbf{X} \leftarrow \mathbf{X}_{\text{tr}}$, use (3) to compute $\boldsymbol{\alpha}_{\text{tr}}$ given $\mathbf{D}$
5: **Compute Test Coefficients:** $\mathbf{X} \leftarrow \mathbf{X}_{\text{ts}}$, use (3) to compute $\boldsymbol{\alpha}_{\text{ts}}$ given $\mathbf{D}$

C. Kernelized Supervised Dictionary Learning

One of the main advantages of the proposed formulation for SDL comparing to other techniques in literature is that we can easily embed a kernel into the formulation. This enables nonlinear transformation of data into a high dimensional feature space where the discrimination of classes can be more efficiently performed. This is especially beneficial by incorporating data dependent kernels[3], such as those based on normalized compression distance [26].

Kernelizing the proposed approach is straightforward. Suppose that $\mathbf{\Psi}$ is a feature map representing the data in feature spaces $\mathcal{H}$ as follows:

$$\begin{aligned}\mathbf{\Psi} : X &\rightarrow \mathcal{H} \\ \mathbf{X} &\mapsto \mathbf{\Psi}(\mathbf{X})\end{aligned} \tag{15}$$

[2]Here, centered data means that the features are centered not individual data samples.

[3]Although it is true that all kernels are computed on the data and hence, are data dependent, the term is used in literature to refer to those types of kernels that do not have any closed form.

To kernelize the proposed SDL, we express the matrix of bases $\mathbf{U}$ as a linear combination of the projected data points into the feature space using representation theory [55], i.e., $\mathbf{U} = \mathbf{\Psi(X)W}$. Replacing $\mathbf{X}$ by $\mathbf{\Psi(X)}$ and $\mathbf{U}$ by $\mathbf{\Psi(X)W}$ into the objective function of (14) we obtain

$$\begin{aligned} \text{tr}(\mathbf{U}^\top\mathbf{\Psi(X)HLH\Psi(X)}^\top\mathbf{U}) &= \text{tr}(\mathbf{W}^\top\mathbf{\Psi(X)}^\top\mathbf{\Psi(X)} \\ &\quad \mathbf{HLH\Psi(X)}^\top\mathbf{\Psi(X)W}) \\ &= \text{tr}(\mathbf{W}^\top\mathbf{KHLHKW}) \end{aligned}$$

with the constraint

$$\begin{aligned} \mathbf{U}^\top\mathbf{U} &= \mathbf{W}^\top\mathbf{\Psi(X)}^\top\mathbf{\Psi(X)W} \\ &= \mathbf{W}^\top\mathbf{KW} \end{aligned}$$

where $\mathbf{K} = \mathbf{\Psi(X)}^\top\mathbf{\Psi(X)}$ is a kernel function on data. Combining this objective function and the constraint, the optimization problem for the kernelized SDL is

$$\begin{aligned} \max_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^\top\mathbf{KHLHKW}) \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{KW} = \mathbf{I} \end{aligned} \tag{16}$$

whose solution is the top eigenvectors of $\mathbf{\Phi} = \mathbf{KHLH}$. Hence, the algorithm for kernelized SDL is given in Algorithm (2).

Algorithm 2 Kernelized Supervised Dictionary Learning

Input: Kernel on training data, $\mathbf{K}_{\text{tr}}$, kernel on test data, $\mathbf{K}_{\text{ts}}$, kernel on labels $\mathbf{L}$, training data size, n, size of dictionary, k.
Output: Dictionary, $\mathbf{D}$, coefficients for training and test data, $\boldsymbol{\alpha}_{\text{tr}}$ and $\boldsymbol{\alpha}_{\text{ts}}$.

1: $\mathbf{H} \leftarrow \mathbf{I} - n^{-1}\mathbf{ee}^\top$
2: $\mathbf{\Phi} \leftarrow \mathbf{KHLH}$
3: **Compute Dictionary:** $\mathbf{D} \leftarrow$ eigenvectors of $\mathbf{\Phi}$ corresponding to top k eigenvalues.
4: **Compute Training Coefficients:** $\mathbf{X} \leftarrow \mathbf{K}_{\text{tr}}$, use (3) to compute $\boldsymbol{\alpha}_{\text{tr}}$ given $\mathbf{D}$
5: **Compute Test Coefficients:** $\mathbf{X} \leftarrow \mathbf{K}_{\text{ts}}$, use (3) to compute $\boldsymbol{\alpha}_{\text{ts}}$ given $\mathbf{D}$

IV. Experiments

In this section, we evaluate the performance of the proposed SDL on various datasets and in different applications such as analyzing face data, digit recognition, classification of real-world data such as satellite images and textures. We will show through various experiments the main advantages of the proposed SDL, such as compact dictionary, i.e., discriminative dictionary even at small dictionary size and fast performance. Also, we will show how its kernelized version enables embedding data dependent kernels into the proposed SDL to significantly improve the performance of difficult classification tasks. Table I provides the details of the datasets used in our experiments along with some details on each dataset including its dimensionality, number of classes, and the number of instances in training and test sets as being used in our experiments.

A. Implementation Details

In our approach, the first step is to compute the dictionary by computing the eigenvectors of $\mathbf{\Phi}$ as provided in Algorithms 1 or 2. To avoid rank deficiency in the computation of kernel on labels, we add identity matrix of the same size to the kernel, i.e., $\mathbf{L} = \mathbf{YY}^\top + \mathbf{I}$. Then we need to calculate the coefficients in the *lasso* provided in (3). We have used the GLMNET[4], which is an efficient implementation of the *lasso* using cyclic coordinate descent [31]. The optimal value of regularization parameter in the *lasso* (λ^*), which controls the level of sparsity, has been computed by 10-fold cross-validation on the training set to minimize the mean-squared error. This λ^* is then used to compute the coefficients for both training and test sets [5].

The same as what is suggested in [56], coefficients computed on the training set are used for training a support vector machine (SVM). RBF kernel has been used for the SVM and the optimal parameters of the SVM, i.e., the optimal kernel width γ^* and trade-off parameter C^* are found by grid search and 5-fold cross-validation on the training set. The coefficients computed on the test set are then submitted to this trained SVM to label unseen test examples. Classification error or accuracy is used to measure the performance of the classification system.

B. Face Data

In this experiment, our main goal is to show the compactness of our proposed dictionary. We use Olivetti face dataset of AT&T [57]. This data is consisting of 400 face images of 40 distinct subjects, i.e., 10 images per subject, at varying lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses). The original size of each image is 92×112 pixels, with 256 gray levels per pixel. However, in our experiments, each image has been cropped from center part to be of size of 64×64 pixels.

The main task in our experiments is to classify the faces to glass/no glass classes. To this end, the images are labeled to indicate these two classes with 119 in glass class and 281 in no glass. Typical images of these two classes are shown in Fig. 1. All images are normalized to have zero mean and unit ℓ_2-norm. Half of the images are randomly selected for training and another half for testing; the experiments are repeated 10 times and the average error is reported in Table 2. The experiments are performed on varying dictionary sizes including 2, 4, 8, 16, and 32. The results are compared with several unsupervised and supervised dictionary learning approaches as shown in Table II. For K-SVD, the fast implementation provided by Rubinstein [58] has been used. We have implemented DK-SVD with K-SVD as the core. The difference between supervised and unsupervised *k*-means is that in unsupervised *k*-means, the dictionary is learned on the

[4]The necessary tools and their Matlab interface can be accessed at http://www-stat.stanford.edu/∼tibs/glmnet-matlab/.

[5]GLMNET handles one data sample at a time and hence one λ^* is computed for each data point in the training set. However, the averaged λ^* over whole training set is used to compute the coefficients on the training and test sets as it yields better generalization.

TABLE I: The datasets used in this paper.

Dataset	Dataset Info.				
	Samples	Training Size	Test Size	Classes	Dim.
Face (Olivetti)[a]	400	200	200	2	4096
Digit (USPS)[b]	9298	7291	2007	10	256
Sonar[c]	208	104	104	2	60
Ionosphere[c]	351	176	175	2	34
Texture (I)[d]	5500	2750	2750	11	40
Satimage[d]	6435	3218	3217	6	36
Texture (II)[e]	600	300	300	2	256

[a]http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
[b]http://www-i6.informatik.rwth-aachen.de/ keysers/usps.html
[c]http://archive.ics.uci.edu/ml/
[d]http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/databases/REAL/
[e]http://www.ux.uis.no/ tranden/

Fig. 1: Typical face images from Olivetti face dataset in two classes of glass vs. no glass.

whole training set whereas in supervised one, one dictionary is learned per class as suggested in texton-based approach by Varma and Zisserman [18], [33]. The code for metaface approach has been provided by the authors [34]. The same as our approach, the parameter(s) of all these rival approaches are tuned using 5-fold cross-validation on the training set.

As can be seen in Table II, our approach performs the best among other approaches. The compactness of the dictionary learned using the proposed SDL is noticeable from the results at small dictionary size. For example, at the dictionary size of two, while the error of our approach is 12.8%, unsupervised k-means yields 27.4% error, which is more than twice as much of the error of our approach. The best result obtained by other supervised dictionary approaches (here metaface) yields 17.55% error at this dictionary size, which is about 5% above the error generated by the proposed SDL. Interestingly, supervised k-means performs significantly better than the unsupervised one particularly at small dictionary sizes. The main conclusion of this experiment is that the proposed SDL generates very discriminative and compact dictionary comparing to well-known unsupervised and supervised dictionary learning approaches.

C. Digit Recognition

The second experiment is performed on the task of handwritten digit classification on the USPS dataset [59]. This dataset is consisting of handwritten digits each with the size of 16×16 pixels with 256 gray levels. There are 7291 and 2009 digits in the training and test sets, respectively.

We compare our results with the most recent SDL technique, which yields the best results published so far on this dataset [42]. To facilitate a direct comparison with what is published in [42], we use the same setup as they have reported. To this end, since the most effective techniques on digit recognition deploy shift invariant features [60], and since neither our approach nor the one reported in [42] benefit from these kind of features, as suggested in [42], the training set is artificially augmented by adding digits which are shifted version of original ones by one pixel to all four directions. Although, this is not an optimal and sophisticated way of introducing shift invariance to the SDL techniques, it takes into account this property in fairly simple approach. Each digit in training and test sets is normalized to have zero mean and unit ℓ_2-norm.

Table III shows the results obtained using the proposed approach in comparison with unsupervised and supervised dictionary learning techniques reported in [42]. As can be seen, again our approach introduces a very compact dictionary such that its performance at dictionary size of 50 is the same as the performance of the system reported in [42] using a dictionary of 100 atoms. With increasing the dictionary size, the performance of our approach slightly degrades. However, it is important to notice that we can achieve a reasonable performance using much less complexity than the best rival. It should be also noted that the best performance achieved by our approach (happening at small dictionary size of 50) is just 0.25% worse than the best results obtained by [42] (happening at dictionary size of 300, i.e., with much higher complexity). This means that our approach misclassifies only 5 more digits compared to the best results obtained in [42] whereas for the same dictionary size (50), our approach performs 0.55% better, i.e., classifies 11 more digits correctly. On the other hand, w.r.t. the complexity, our proposed approach offers a much simpler solution for SDL than the approach in [42]: there are fewer parameters to tune, the dictionary can be computed in closed form, and there is no need to solve a complicated noncovex optimization problem as is used in [42] by iteratively and alternatively optimizing classifier, dictionary, and coefficients learning.

As the final remark, due to orthonormality constraint in the optimization problem of our proposed SDL as given in (14), overcompleteness is not possible in our proposed SDL. This is the reason that in Table III, no results are reported for dictionary size of 300 for our approach. However, as

TABLE II: Classification error on test set for Olivetti face data using the proposed SDL. The results are compared with several other dictionary learning approaches in the literature. The best results obtained are highlighted.

Approach		Dictionary Size				
		2	4	8	16	32
Unsupervised	k-means	27.40 $\pm$2.04	22.60 $\pm$5.18	13.15 $\pm$2.38	8.15 $\pm$1.81	5.75 $\pm$1.70
	K-SVD [29]	28.20 $\pm$2.45	20.60 $\pm$2.41	9.65 $\pm$1.62	7.75 $\pm$2.06	4.05 $\pm$1.23
Supervised	Proposed SDL	**12.80 $\pm$3.77**	**10.05 $\pm$3.11**	**4.95 $\pm$1.92**	**4.95 $\pm$1.14**	**3.30 $\pm$1.53**
	DK-SVD [39]	17.80 $\pm$3.06	10.25 $\pm$2.48	8.75 $\pm$2.02	7.05 $\pm$2.11	6.75 $\pm$1.53
	k-means[a] [18]	17.75 $\pm$3.65	10.40 $\pm$2.56	7.40 $\pm$1.90	5.55 $\pm$1.62	3.65 $\pm$1.20
	Metaface [34]	17.55 $\pm$2.87	11.25 $\pm$2.35	9.75 $\pm$3.58	7.60 $\pm$1.39	5.45 $\pm$0.96

[a]Supervised k-means learns one sub-dictionary per class and then compose all learned sub-dictionaries into one.

TABLE III: Classification error on test set for digit recognition on USPS data using proposed SDL compared with the most effective SDL approach reported in the literature on the same data [42]. Highlighted entries represent the best results obtained at each dictionary size.

Approach	Dictionary Size			
	50	100	200	300
Unsupervised [42]	8.02	6.03	5.13	4.58
Supervised [42]	3.64	**3.09**	**2.88**	**2.84**
Proposed SDL	**3.09**	3.19	3.64	-

mentioned above, due to the compactness of our dictionary, good results are obtained at much smaller dictionary size, which is a desired attribute as it decreases the computational load. Also, the proposed kernelized version of our proposed approach given in (16) and Algorithm 2, can learn dictionaries as large as n, i.e., the number of data points used for training, which is usually greater than the dimensionality of the data p (see Table I for the relative size of p and n for the data used in our experiments).

D. Other Real-World Data

In two previous sections, the classification task was performed on the pixels of images directly. In this section, we evaluate the performance of the proposed approach on the classification of some real-world data using features extracted. Four datasets with varying complexity from 2- to 11-class, with the dimensionality of up to 60 features, and also with as many as 6435 data samples are used in these experiments (refer to Table I for detailed information on these datasets). All data are preprocessed to have zero mean and unit ℓ_2-norm, except Satimage dataset, where the features are normalized to be in the range of [0, 1] due to the large variation of feature values.

On all datasets, the experiments are repeated 10 times over random split of data into half for training and another half for testing. The average and standard deviation of classification accuracy are reported in Table IV in comparison with several other unsupervised and supervised dictionary learning approaches. We have also included the results of classification using kernelized version of our proposed SDL with radial basis function (RBF) as kernel. The width of the RBF kernel has been selected based on self-tuning approach [61].

As can be seen from Table IV, the proposed SDL or its kernelized version performs the best in all cases except for the dictionary size of 8 and 16 on Sonar data. DK-SVD performs poorly (even worse than unsupervised K-SVD approach) on these datasets mainly because, by design, it uses a linear classifier (refer to Subsection II-B and [39] for more description on this approach). The poor performance of metaface is because it usually performs well at very large dictionary size. Hence, at reported dictionary sizes, its training is not sufficient to capture the underlying data structure. For example, for Sonar data, while proposed SDL can achieve the accuracy of 79.23$\pm$4.67 at the dictionary size of 32, metaface approach can only achieve this accuracy at the dictionary size of 64 (accuracy 80.00$\pm$4.75). However, using large dictionary size adds to the computational load of the approach.

E. Patch Classification on Texture Data

To show the benefit from using data-dependent kernels such as kernels computed using normalized compression distance [26], in this section, we perform classification on patches extracted from texture images. We compare our results with and without kernels on the proposed approach and also compare them to the results published in [15], i.e., two supervised dictionary learning approaches called SDL-G BL(G for generative and BL for bilinear model) and SDL-D BL(D for discriminative). To ease the comparison, we use the same data as in [15], i.e., classification on texture pair of D5 and D92 from Brodatz album shown in Fig. 2. Also the same as [15], 300 patches are randomly extracted from the left half of each texture image for training and 300 patches from right half for testing. This is to ensure that there is no overlap among the patches used in the training and test sets.

We have used RBF kernel and two data-dependent compression-based kernels as reported in [62] (CK-1) and [27] (d_{N}) as the kernel for the proposed kernelized SDL. The latter deploys MPEG-1 as the compressor as suggested in [62] for the computation of normalized compression distance [26].

TABLE IV: The results of classification accuracy (%) on various real-world datasets using different methods and in different dictionary sizes. The best results obtained are highlighted.

Approach		Sonar			Ionosphere			Texture			Satimage		
		8	16	32	8	16	32	8	16	32	8	16	32
Unsupervised	k-means	71.44	75.48	75.58	92.63	92.29	91.94	97.51	98.88	99.03	86.64	86.98	87.13
		±5.53	±5.43	±3.77	±2.48	±1.41	±1.86	±0.66	±0.25	±0.29	±0.47	±0.64	±0.72
	K-SVD [29]	72.69	75.19	71.44	91.31	90.91	92.00	98.46	99.19	99.17	89.58	89.30	88.08
		±2.69	±6.69	±4.25	±4.12	±1.73	±1.50	±0.30	±0.27	±0.19	±0.43	±0.73	±0.36
Supervised	Proposed SDL	72.21	77.50	**79.23**	**94.06**	**94.40**	**94.57**	**98.56**	**99.55**	**99.69**	88.75	89.42	89.34
		±3.47	±2.73	**±4.67**	**±1.66**	**±1.41**	**±1.41**	**±0.38**	**±0.12**	**±0.10**	±0.36	±0.40	±0.41
	KSDL-RBF[a]	74.81	75.67	75.96	94.17	94.06	94.11	98.47	99.22	99.25	**90.05**	**90.61**	**90.59**
		±4.00	±4.00	±5.53	±1.82	±1.91	±2.05	±0.13	±0.08	±0.11	**±0.43**	**±0.39**	**±0.42**
	DK-SVD [39]	67.60	67.31	70.96	83.89	82.00	84.11	72.09	93.85	92.72	64.64	79.85	71.11
		±4.53	±4.32	±4.15	±1.88	±3.51	±2.50	±3.87	±0.82	±1.86	±13.29	±1.38	±4.19
	k-means [18]	**75.38**	**77.69**	77.12	92.46	90.46	90.00	97.89	99.05	99.18	86.39	87.35	87.02
		±5.31	**±4.27**	±5.98	±1.39	±1.59	±2.35	±0.42	±0.14	±0.22	±0.36	±0.32	±0.65
	Metaface [34]	73.26	72.11	76.83	81.71	78.46	83.71	90.24	89.97	95.36	76.57	72.86	75.15
		±3.17	±5.22	±4.43	±1.62	±2.89	±2.52	±0.55	±1.88	±0.57	±1.38	±1.05	±1.53

[a]Proposed kernel SDL with RBF kernel.

However, comparing to the measure proposed in [62] (CK-1), it proposes a novel compression-based dissimilarity measure (d_{N}) that performs well on both small and large patch sizes (as shown in [27], CK-1 does not work properly on small patch sizes). Besides, d_{N} is a semi-metric.

Table V provides the results of classification using proposed SDL with and without kernels. It also compares the results with k-means as an unsupervised approach to compute the dictionary and also with the results published in [15] for the same number of patches (300). The sparsity of the coefficients (i.e., the number of nonzero coefficients) are also provided in this table (it is not reported for SDL-G BL and SDL-D BL in [15]). As can be seen, using compression-based data dependent kernel based on d_{N} dramatically improves the results. The classification error is even lower than the one obtained by SDL-D BL approach using 30000 patches for training, which yields the best results on this data in [15] (classification error = 14.26%). Moreover, as the sparsity of the coefficients indicate, the proposed approach with data-dependent kernel d_{N} deploys smallest number of dictionary atoms in the reconstruction of signal, i.e., benefits the most from the sparse representation (it almost uses half of the dictionary elements comparing to other approaches). This has a great impact on the computation load of the classification task especially in the stage of training and testing of the classifier. Our experiments show (not reported in Table V) by using a slightly larger regularization parameter λ in the *lasso* such that the reconstruction error is within one standard deviation of the minimum, the sparsity of coefficients can be even more (about one third of coefficients are nonzero) without compromising the classification error (the classification error is 10.52 ± 1.38 in this case, which is not very much different from what is reported in Table V).

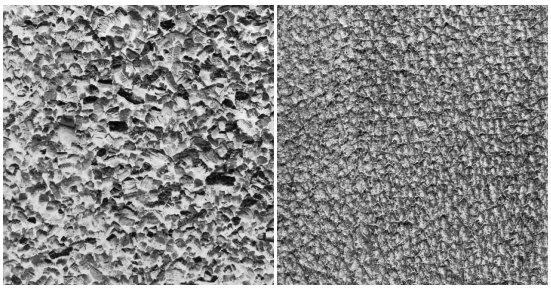

Fig. 2: Texture images of D5 and D92 from Brodatz album.

V. Discussions and Conclusions

In this paper we proposed a novel supervised dictionary learning. The proposed approach learns the dictionary in a space where the dependency between the data and category information is maximized. Maximizing this dependency has been performed based on the concept of Hilbert Schmidt independence criterion (HSIC). This introduces a data decomposition that represents the data in a space with maximum dependency with category information. We showed that the dictionary can be learned in this space in closed form. The sparse coefficients can be learned by using the *lasso* as given in (3). Our experiments using real-world data with varying complexity shows that the proposed approach is very efficient in classification tasks and outperforms other unsupervised and supervised dictionary learning approaches in the literature. Besides, the proposed approach is very fast and efficient in computation.

We also showed how the proposed SDL can be kernelized. This enables the proposed SDL to benefit from data dependent kernels. It was shown using some experiments that proposed kernelized SDL can significantly improve the results in difficult classification tasks comparing to other SDL approaches in the literature. To our best of knowledge, this is the first SDL in the literature that can be kernelized and benefit from data-dependent kernels embedded into the SDL.

The proposed approach learns a very compact dictionary in the sense that it significantly outperforms other approaches when the size of dictionary is very small. This shows that the proposed SDL can effectively encode the category information into the learning of dictionary such that it can perform very well in classification tasks using few atoms.

In dictionary learning literature, usually the dictionary

TABLE V: Classification error and the number of nonzero coefficients on the test set for texture pair D5-D92 of Brodatz album. Using data-dependent kernels and proposed kernelized SDL can significantly improve the results.

Approach		Average No. of Nonzero Coefficients		Classification Error (%)
		Train Set	Test Set	
k-means		47.85	48.99	27.75±2.29
Proposed SDL		59.80	59.85	26.43±2.95
Proposed kernel SDL	RBF	62.86	62.30	30.13±2.81
	CK-1 [62]	58.76	58.63	22.15±1.34
	d_{N} [27]	**34.03**	**31.72**	**10.37±1.37**
SDL-G BL [15]		-	-	26.34
SDL-D BL [15]		-	-	26.34

learned is overcomplete, i.e., the number of elements in the learned dictionary is larger than the dimensionality of the data/dictionary. In our proposed SDL, due to orthonormality constraint on the dictionary atoms, as can be seen in (14), the dictionary cannot be overcomplete. However, there are two remarks here: First, as discussed above, our dictionary is very compact and as the experiments show, the proposed SDL performs very well at small dictionary size, which is usually below even complete dictionary size. This is a main advantage of the proposed approach as small dictionary size means lower computational cost. Second, the kernelized version of the proposed approach can easily learn dictionaries as large as n, the number of data samples in the training set. This is because the kernel computed on the data is of the dimensionality of n, which is usually greater than p (the dimensionality of data). Note that for all datasets provided in this paper except Olivetti face dataset, the number of data in training set is larger than the dimensionality of data (refer to Table I). For face dataset, it is worth to note that a dictionary as small as 32 atoms leads to extremely good results using proposed SDL and overcompleteness is not necessary here.

Another advantage of proposed approach is that there is only one parameter to be tuned, which is the regularization parameter λ in the *lasso*. Since the dictionary is learned in closed form, it is extremely fast to tune this parameter within the classification task or by minimizing the reconstruction error. This is while there are usually several parameters in other SDL approaches in the literature to be tuned and since learning the dictionary and coefficients have to be performed alternatively and iteratively, it is very time consuming to tune these parameters using a cross validation on the training set.

In this research, we have used a SVM with RBF kernel on the sparse coefficients learned for performing the classification task. This may not fully utilize the sparsity. In future work, we will consider other kernels for the SVM or other classifiers that can benefit more from sparse nature of data points submitted for classification as suggested in [56].

Also, we proposed to use $\mathbf{L} = \mathbf{Y}\mathbf{Y}^\top + \mathbf{I}$ as the kernel on the labels. As proposed in [63], [64], it is possible to encode the relationship among the classes into a matrix $\mathbf{M} \in \mathbb{R}^{c \times c}$, where c is the number of classes, and use $\mathbf{L} = \mathbf{Y}\mathbf{M}\mathbf{Y}^\top + \mathbf{I}$ instead to build up the kernel on the labels. This may consequently better encode the data structure into the learning of dictionary and as a future work, we will implement this in the formulation provided for Algorithm 1.

REFERENCES

[1] S. Mallat, *A Wavelet Tour of signal Processing: The Sparse Way*, 3rd ed. Academic Press, 2009.

[2] B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Mar. 1996.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: John Wiley & Sons, 2001.

[4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[6] M. Biggs, A. Ghodsi, and S. Vavasis, "Nonnegative matrix factorization via rank-one downdate," in *Proceedings of the* 25th *International Conference on Machine Learning (ICML)*, 2008, pp. 64–71.

[7] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[8] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[9] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Science Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[10] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in 27th *Asilomar Conference on Signals, Systems and Computers*, 1993, pp. 40–44.

[11] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[12] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1031–1044, June 2010.

[13] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[14] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York: Springer, 2010.

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 1033–1040.

[16] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, "Dictionary learning in texture classification," in *Proceedings of the* 8th *international conference on Image analysis and recognition - Volume Part I*. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 335–343.

[17] J. Xie, L. Zhang, J. You, and D. Zhang, "Texture classification via patch-based sparse texton learning," in 17th *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 2737–2740.

[18] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2032–2047, Nov. 2009.

[19] C. Zhong, Z. Sun, and T. Tan, "Robust 3D face recognition using learned visual codebook," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–6.

[20] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[21] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in 13th *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 543–550.

[22] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[23] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, Jan. 2008.

[24] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Proceedings of the* 16th *international conference on Algorithmic Learning Theory (ALT)*, 2005, pp. 63–77.

[25] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf, "Kernel methods for measuring independence," *Journal of Machine Learning Research*, vol. 6, pp. 2075–2129, Dec. 2005.

[26] R. Cilibrasi and P. Vitányi, "Clustering by compression," *IEEE Trans. Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.

[27] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, "Supervised texture classification using a novel compression-based similarity measure," in *Proceedings of the International Conference on Computer Vision and Graphics (ICCVG)*, 2012.

[28] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 609–616.

[29] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[30] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal Machine Learning Research*, vol. 11, pp. 19–60, Mar. 2010.

[31] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, Feb. 2010.

[32] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, pp. 29–44, June 2001.

[33] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis*, vol. 62, no. 1-2, pp. 61–81, 2005.

[34] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Metaface learning for sparse representation based face recognition," in 17th *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1601–1604.

[35] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3501–3508.

[36] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proceedings of the* 10th *European Conference on Computer Vision (ECCV): Part I*, 2008, pp. 179–192.

[37] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in 10th *IEEE International Conference on Computer Vision (ICCV)*, 2005, pp. 1800–1807.

[38] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[39] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2691–2698.

[40] D.-S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[41] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[42] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[43] S. Lazebnik and M. Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1294–1309, July 2009.

[44] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 985–992.

[45] F. Rodriguez and G. Sapiro, "Sparse representations for image classification: learning discriminative and reconstructive non-parametric dictionaries," in *IMA Preprint Series 2213*, 2007.

[46] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[47] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: John Wiley & Sons, 2006.

[48] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[49] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.

[50] L. Song, J. Bedo, K. Borgwardt, A. Gretton, and A. Smola, "Gene selection via the bahsic family of algorithms," *Bioinformatics*, vol. 23, pp. i490–i498, July 2007.

[51] H. Shen, S. Jegelka, and A. Gretton, "Fast kernel-based independent component analysis," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3498–3511, Sept. 2009.

[52] N. Quadrianto, A. Smola, L. Song, and T. Tuytelaars, "Kernelized sorting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1809–1821, Oct. 2010.

[53] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample problem," Max Planck Institute for Biological Cybernetics, Technical Report 157, Apr. 2008.

[54] S. Jegelka, A. Gretton, B. Schölkopf, B. K. Sriperumbudur, and U. Von Luxburg, "Generalized clustering via kernel embeddings," in *Proceedings of the* 32nd *Annual German Conference on Advances in Artificial Intelligence*, 2009, pp. 144–152.

[55] J. L. Alperin, *Local Representation Theory: Modular Representations as an Introduction to the Local Representation Theory of Finite Groups*. New York: Cambridge University Press, 1986.

[56] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the* 24th *international conference on Machine learning (ICML)*, 2007, pp. 759–766.

[57] *Cambridge University Computer Laboratory, Olivetti Face Dataset AT&T*, 1994, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.

[58] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," Dept. of Computer Science, Technion, Technical Report, 2008.

[59] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1990, pp. 396–404.

[60] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[61] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems (NIPS)*, 2004, pp. 1601–1608.

[62] B. Campana and E. Keogh, "A compression-based distance measure for texture," *Statistical Analysis and Data Mining*, vol. 3, no. 6, pp. 381–398, 2010.

[63] M. Blaschko and A. Gretton, "Taxonomy inference using kernel dependence measures," Max Planck Institute for Biological Cybernetics, Technical Report 181, Nov. 2008.

[64] L. Song, A. Smola, A. Gretton, and K. Borgwardt, "A dependence maximization view of clustering," in *Proceedings of the* 24th *international conference on Machine learning (ICML)*, 2007, pp. 815–822.